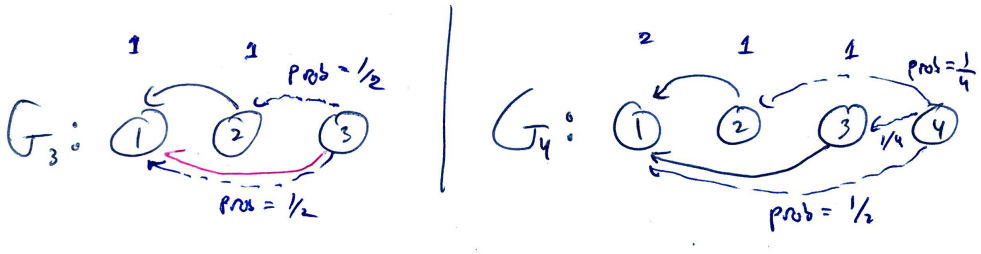
Lecture 29: Preferential attachment tree

Thus far we have been focusing on the Erdős-Rényi random graph, which is mathematically interesting, but is not such a realistic model for real-world networks. Just intuitively, the Erdős-Rényi random graph has too much independence for many settings: in a social network, if two people (vertices) are friends (connected by an edge), they are more likely to have friends in common (both be connected to the same vertex). But more concretely, the distribution of various statistics don't match up with real-world networks. In particular, we saw in Lecture 9 that the degree distribution of an Erdős-Rényi random graph with average degree λ has a degree distribution that is close to Poisson. A Poisson distribution has exponentially decaying tails (Lecture 16), while it has been observed that many real world networks have a degree distribution that decays like a power; this is called "power law" behaviour (see Chapter 1 of van der Hofstad for a more detailed high-level discussion on this point and others). This means the proportion of vertices with degree k is of order $k^{-\tau}$ for large k and some $\tau > 1$ (the inequality because the tail must be summable). There are a number of random graph models that give power law behaviour, and for the next few lectures we will study a particular one that was popularised in the early 2000's, and was a seminal contribution to "network science".

Preferential attachment random graphs. A preferential attachment random graph is generated sequentially, by adding vertices one at a time, with each vertex attaching edges to the existing structure. The chance that an existing vertex receives an edge from an arriving vertex increases with the degree of the existing vertex; hence the name preferential attachment. There are a huge number of preferential attachment models that differ in the details of this construction, such as the initial graph to start the process, the increasing function determining the probabilities of connection, the number of edges an arriving vertex attaches to the existing structure, etc. Here we study the simplest model to keep the computations simple, but after understanding this case, the more general model in Chapter 8 of van der Hofstad is very approachable.

Definition 29.1 (Preferential attachment graph)**.** We define a sequence of random graphs (trees) $(G_n)_{n\geq 2}$ where G_n has n vertices and $n-1$ edges. The initial graph G_2 consists of two vertices labeled 1 and 2 with a single directed edge from vertex 2 to vertex 1. Given G_n, the graph G_{n+1} is obtained by adding a vertex labeled $(n+1)$ and a single directed edge from this new vertex to a vertex labeled from the set $\{1,\dots,n\}$, where the chance that vertex $(n+1)$ connects to vertex i is proportional to the degree of vertex i in G_n (here and below degree means in-degree plus out-degree). We denote the law of the preferential attachment graph on n vertices by PA_n. See the picture for illustration.



Our main result, which we will prove over the next few lectures is the following.

Theorem 29.2. *Let $G_n \sim \mathrm{PA}_n$, $D_n(i)$ be the degree of vertex i in G_n, and*

$$P_n(k) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[D_n(i) = k]$$

be the proportion of vertices having degree k in G_n. Setting $p_k := \frac{4}{k(k+1)(k+2)}$, there is a constant c such that

$$\mathbb{P}\Big(\max_{k\geq 1}\big|P_n(k) - p_k\big| \geq c\sqrt{\log(n)/n}\Big) \xrightarrow[n\to\infty]{} 0.$$

The theorem shows that the asymptotic empirical degree distribution is close to the probability distribution with mass function $(p_k)_{k\geq 1}$ (it's an easy exercise to see this is a probability distribution). For large k, the probabilities satisfy $p_k \asymp k^{-3}$, and this is the power law behaviour (with exponent 3) previously mentioned. By giving the vertices initial "weights", and having arriving vertices attach more than one edge, it is possible to obtain power law behaviour with other exponents (rather than 3); see Chapter 8 of van der Hofstad.

We proved an analogous result for the Erdős-Rényi random graph in Lecture 9 (Proposition 9.1). The proof of Theorem 29.2 follows a similar strategy, we first show $P_n(k)$ is concentrated around its mean, and then show its mean is close to p_k. Recalling the discussion there, Theorem 29.2 shows that the maximum deviation of the $P_n(k)$ from their mean is of order at most $\sqrt{\log(n)/n}$, as opposed to smaller order than $\omega(1/\sqrt{n})$ for the Erdős-Rényi random graph.

We state the concentration and mean approximation results in the following lemma.

Lemma 29.3. *Recall the notation and definitions of Theorem 29.2 just above.*

1. *There is a constant $c > 0$ such that*

$$\mathbb{P}\Big(\max_{k\geq 1}\big|P_n(k) - \mathbb{E}\big[P_n(k)\big]\big| \geq c\sqrt{\log(n)/n}\Big) \xrightarrow[n\to\infty]{} 0.$$

2. *There is a constant $c > 0$ such that for all $k \geq 1$,*

$$\big|\mathbb{E}\big[P_n(k)\big] - p_k\big| \leq \frac{c}{n}.$$

The theorem easily follows from this lemma in the same way as in the proof of Proposition 9.1; see (9.2), so all that is left to do is to prove the lemma. For the first part of the lemma, we need to develop an exponential tail bound inequality. For the second part, we condition on G_{n-1} to write $\mathbb{E}[P_n(k)]$ in terms of $\mathbb{E}[P_{n-1}(k)]$ and $\mathbb{E}[P_{n-1}(k-1)]$, and then use induction on n and k with this recursion.

Tool: Azuma-Hoeffding inequality. First we need the Azuma-Hoeffding inequality.

Lemma 29.4. *Let $(M_i)_{i=0}^n$ be random variables, and for $i = 0, \ldots, n$, let $(\mathcal{F}_i)_{i=0}^n$ be an increasing sequence of sigma-algebras such that $\mathcal{F}_0$ is trivial (the empty set and the whole space) and $\sigma(M_1, \ldots, M_i) \subseteq \mathcal{F}_i$. If for $i = 1, \ldots, n$, $\mathbb{E}\big[|M_i|\big] < \infty$, and*

$$\mathbb{E}[M_i|\mathcal{F}_{i-1}] = M_{i-1}, \tag{29.1}$$

and

$$|M_i - M_{i-1}| \leq a_i,$$

then for any $t > 0$,

$$\mathbb{P}\big(|M_n - M_0| \geq t\big) \leq 2e^{-t^2/(2\sum_{i=1}^n a_i^2)}.$$

Proof. We use the Chernoff-Cramér method. Write

$$M_n - M_0 = \sum_{i=1}^n (M_i - M_{i-1}),$$

and note that for any $\theta \in \mathbb{R}$,

$$\mathbb{E}\big[e^{\theta(M_n - M_0)}\big] = \mathbb{E}\left[\prod_{i=1}^n e^{\theta(M_i - M_{i-1})}\right]$$

$$= \mathbb{E}\Big\{ \mathbb{E}\Big[e^{\theta(M_n - M_{n-1})}\Big|\mathcal{F}_{n-1}\Big] \prod_{i=1}^{n-1} e^{\theta(M_i - M_{i-1})} \Big\}$$

$$\leq e^{\theta^2 a_n^2/2} \mathbb{E}\Big\{ \prod_{i=1}^{n-1} e^{\theta(M_i - M_{i-1})} \Big\}, \tag{29.2}$$

where in the inequality we have used Exercise 1(a) from Assignment 2, and note that here is where we are using (29.1), so that

$$\mathbb{E}[M_n - M_{n-1}|\mathcal{F}_{n-1}] = 0.$$

Iterating this argument starting from the expectation in (29.2) implies that

$$\mathbb{E}\big[e^{\theta(M_n - M_0)}\big] \leq e^{\theta^2 \sum_{i=1}^n a_i^2/2}. \tag{29.3}$$

Now note

$$\mathbb{P}(|M_n - M_0| \geq t) = \mathbb{P}(M_n - M_0 \geq t) + \mathbb{P}(M_n - M_0 \leq -t).$$

Both terms are bounded in the same way. For the second term, we use the Chernoff-Crámer method to find for $\theta < 0$ that

$$\mathbb{P}\big(M_n - M_0 \leq -t\big) = \mathbb{P}\big(e^{\theta(M_n - M_0)} \geq e^{-\theta t}\big) \leq \mathbb{E}[e^{\theta(M_n - M_0)}]e^{\theta t} \leq e^{\theta t} e^{\theta^2 \sum_{i=1}^n a_i^2/2},$$

where in the inequality we use (29.3). Now choosing $\theta = -t/\sum_{i=1}^n a_i^2 < 0$ gives the result. □

Sequences $(M_i)_{i=0}^n$ satisfying condition (29.1) are called *martingales* with respect to the *filtration* $(\mathcal{F}_i)_{i=0}^n$ and they have many nice properties. These are covered in other subjects, so we don't dwell on the general structure here, and instead show the following corollary in the case where $(M_i)_{i=0}^n$ is a *Doob martingale*, which is stated in way that is specialised to our setting.

Lemma 29.5. *Let $(G_1, \ldots, G_n)$ be random elements such that G_{i-1} is measurable with respect to $\sigma(G_i)$ (or G_{i-1} is a function of G_i), and let X be a random variable on the same space such that $\mathbb{E}\big[|X|\big] < \infty$ and for $i = 1, \ldots, n$,*

$$\big|\mathbb{E}[X|G_i] - \mathbb{E}[X|G_{i-1}]\big| \leq a_i.$$

then for any $t > 0$,

$$\mathbb{P}\big(|\mathbb{E}[X|G_n] - \mathbb{E}[X]| \geq t\big) \leq 2e^{-t^2/(2\sum_{i=1}^n a_i^2)}.$$

Proof. We use the Azuma-Hoeffding inequality of Lemma 29.4 with $M_i := \mathbb{E}[X|G_i]$ and $\mathcal{F}_i = \sigma(G_i)$. we have

$$\mathbb{E}\big[|M_i|\big] = \mathbb{E}\big[|\mathbb{E}[X|G_i]|\big] \leq \mathbb{E}\big[\mathbb{E}[|X||G_i]\big] = \mathbb{E}\big[|X|\big] < \infty,$$

and

$$\mathbb{E}[M_i|\mathcal{F}_{i-1}] = \mathbb{E}\big[\mathbb{E}[X|G_i]\big|G_{i-1}\big] = \mathbb{E}[X|G_{i-1}] = M_{i-1},$$

where the second equality is because G_{i-1} is measurable with respect to $\sigma(G_i)$. □

These last two lemmas may look a bit abstract if you have not seen many examples of martingales. In the last lemma, the notation is purposefully suggestive: You can think of the G_i as the preferential attachment graph with i vertices, generated sequentially. We can then take X to be $P_n(k)$, which is G_n measurable, so $\mathbb{E}[P_n(k)|G_n] = P_n(k)$. Then we will see that the difference between $\mathbb{E}[P_n(k)|G_i]$ and $\mathbb{E}[P_n(k)|G_{i-1}]$ is small, essentially because we can control the differences between these values through what happens at step i.